DOCUMENT RESUME

ED 271 497 TM 860 406

AUTHOR Mandeville, Garrett K.; Anderson, Lorin W.
TITLE A Study of the Stability of School Effectiveness Measures across Grades and Subject Areas.
PUB DATE Apr 86
NOTE 23p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 16-20, 1986).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Academic Achievement; Achievement Rating; Achievement Tests; *Age Differences; Elementary Education; *Elementary School Mathematics; Elementary Schools; *Evaluation Criteria; Evaluation Methods; *Reading Achievement; Regression (Statistics); Research Methodology; *Research Problems; *School Effectiveness; Scores; Socioeconomic Influences; State Programs; Testing Programs; Test Reliability
IDENTIFIERS Cognitive Skills Assessment Battery; Comprehensive Tests of Basic Skills; *Effective Schools Research; South Carolina Basic Skills Assessment Program

ABSTRACT

School effectiveness indices (SEIs), based on regressing test performance onto earlier test performance and a socioeconomic status measure, were obtained for eight subject-grade combinations from 485 South Carolina elementary schools. The analysis involved school means based on longitudinally matched student data. Reading and mathematics achievement data were gathered from the South Carolina Basic Skills Assessment Program tests, the Comprehensive Tests of Basic Skills, and the Cognitive Skills Assessment Battery. Grades one through four were included. The resulting SEIs were found to be somewhat unstable across subject areas and very unstable across grades. Grade-to-grade correlations of the SEIs measuring mathematics performance, although small, were largely significant whereas those measuring reading performance were generally nonsignificant. This suggested that school effects may be more readily discernible in some subject areas than in others. Implications were drawn for effective schools research and for school incentive award systems based on student test performance. (Author/GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED271497

A Study of the Stability of School Effectiveness Measures Across Grades and Subject Areas

Garrett K. Mandeville
and
Lorin W. Anderson
University of South Carolina

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
G. K. Mandeville
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
X This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

BEST COPY AVAILABLE

TM 860 406

Paper presented at the annual meeting of the
National Council on Measurement in Education
San Francisco, April 1986.

Abstract

School effectiveness indices (SEIs) based on regressing test performance onto earlier test performance and an SES measure, were obtained for eight subject-grade combinations for a large sample of elementary schools. The analyses involved school means based on longitudinally matched student data. The resulting SEIs were found to be somewhat unstable across subject areas (reading and mathematics) and very unstable across grades (one through four). Grade-to-grade correlations of the SEIs measuring mathematics performance, although small, were largely significant whereas those measuring reading performance were generally nonsignificant. This suggested that school effects may be more readily discernible in some subject areas than in others. Implications were drawn for research on effective schools and school incentive award systems based on student test performance.

For a number of years, researchers have been attempting to examine how well individual schools have been doing in their efforts to foster important educational outcomes in the children who attend them. Most frequently this examination has utilized quantitative indicators of overall student performance and the focus has been on school accountability, "school effectiveness", and the more recent efforts to award schools whose students have exhibited exceptional achievement. (None of these movements should be confused with the estimation of what have been called "school effects" (e.g., Coleman, et al.,1966) in which the objective has been to estimate the portion of achievement variation which can be attributed to schools in general after various background factors have been taken into account.)

The recent impetus in this area is related to state- and district-level programs to monitor school performance on the basis of student achievement test data. In some cases, these programs lead to recognition of high performing schools and in a few cases, monetary awards to the schools and/or their personnel (see, e. g., Wynne, 1984). At the state level California, Florida, and South Carolina now have school award programs in which test scores are a major factor in the determination of awardees. District-level programs would include the Dallas Independent School District and the Montgomery County (MD) Public Schools to name but a few. In some cases district-level programs derive from the "effective schools" literature, the objective being to identify and then study high performing schools rather than simply to reward them in some way.

This interest has led to a number of research papers dealing with methodological problems in the identification of schools to receive recognition based on student achievement. Although these papers have rather diverse objectives, they tend to concentrate on

comparing the results from utilizing the various methodologies which have been proposed. When similar methodologies have been compared (e. g., assorted regression approaches) the results tend to be quite consistent (e. g., Webster and Olson, 1984; Abalos, Jolly, and Johnson, 1985), but when the procedures vary on major dimensions, the reverse is usually true (e.g., Frechtling, 1982; Frederick and Clauset, 1985). In these latter situations, researcher's recommendations concerning which procedure to use have often been based on equity issues such as lack of bias toward low SES, under-achieving, or minority children.

In much of the earlier "effective schools" research, schools were identified based on the performance of a rather limited sample of their student body (e.g., students at one grade level for one year) and in some cases the performance of these students in only one subject area (e.g., reading) was considered. Although other methodological problems received more attention, critics such as Rowan, Bossert and Dwyer (1983) and Ralph and Fennessey (1983) have also taken these researchers to task for the rather limited nature of many of these earlier studies.

In some cases, researchers of "school award" algorithms, have also been guilty of limiting the purview of their analyses, but the trend is toward computing two or more indices at each grade level and then aggregating these indices to the school level. Although the issue of how best to conduct this aggregation is beginning to receive some attention (Abalos, Jolly, and Johnson, 1985), researchers do not seem to be concerned about whether this aggregation is sensible from a psychometric standpoint.

This paper will investigate the consistency of what will be termed "school effectiveness indices" (SEI) as a function of grade

level and subject matter. This issue is worthy of study for two reasons. First, in some cases, the criticisms noted above concerning limited grade level and subject matter coverage are still relevant. Second, it addresses an assumption implicit in aggregation, namely that comparable indices are being aggregated. SEIs will be constructed for the two basic skills areas of reading and mathematics for each of grades one through four for a large sample of elementary schools and the consistency of the resultant indices will be considered. Implications for the identification of "effective schools" and those to receive awards will be discussed.

Theoretical and Empirical Background

Rowan, Bossart and Dwyer (1983) have identified the following four general approaches to the creation of SEIs: (1) the use of absolute standards such as school means and comparing them to national norms, (2) analyzing trends in test scores for a given grade level over a period of years, (3) analyzing trends in test scores for a given cohort of students as they progress through a school, possibly comparing their performance to national normative data, and (4) various methods based on residuals from a regression analysis. As noted above, the results from applying various somewhat distinct approaches have been demonstrated to be quite inconsistent. The family of approaches which appears to have the most empirical support, however, are those based on regressing achievement onto prior achievement and some measure of socio-economic-status (SES). For these reasons, this was the general approach selected for this investigation.

An early, rather well-known regression-based methodology was developed by Dyer and his colleagues (Dyer, Linn and Patton, 1969; Dyer, 1970). Basically an educational accountability system, in the

"student change model of an educational system" student outcomes are regressed onto student inputs and "hard-to-change" variables, with the residuals serving as the SEIs (or as the basis of what Dyer called "performance indicators" (PI)). The PI metric was simply a five point scale based on the standardized residuals. One important characteristic of Dyer's system was the use of longitudinal data, the justification being that, "the only fair index of school effectiveness is one that rests on input-output data obtained only on those pupils with whom the school staff has been in continuous contact over a specified period of months or years." (Dyer, 1970, p. 208). Results presented by Hilton and Patrick (1970) demonstrated differences between school aggregate indices based on matched longitudinal, unmatched longitudinal, and cross-sectional data. Related results in Dyer, et al. (1969) also provided empirical support for this position.

A related issue involves the unit of analysis to use in the regression analysis. Dyer, et al. (1969) found that the residuals from an individual level regression analysis aggregated to the school level were highly correlated (median r=.93) with the residuals from an analysis involving school means. However, they also found that the individual level analysis produced SEIs which were slightly more stable. O'Connor (1972), however, noted that the aggregation of the individual level residuals to the school level produces summary values which are correlated with both inputs and predicted outputs since the individual level regression coefficients do not provide the least squares solution to the problem of interest. For these reasons and the empirical support provided by Frechtling (1982), regression of school means was selected as the analytic strategy for this investigation.

As noted above, Dyer et al. (1969) provided some assessment of

the stability of the SEIs generated from the four regression strategies they considered. The study involved 64 school systems (rather than schools) with standardized achievement test results for eighth graders in the 1960-61 school year being regressed onto the corresponding scores of these students when they were fifth graders. The sample from each school system was split into two random subsamples of equal size and, for the analysis involving school means, the intercorrelations between the pairs of SEIs ranged from .62 to .84 depending on the subject area being tested. When these correlations were stepped-up using the Spearman-Brown formula to reflect the reliability of the composite (a more appropriate index of the stability for total sample; see O'Connor, 1972) the coefficients ranged from .77 to .91.

In a similar investigation Marco (1974) studied a sample of third grade Title I students enrolled in 70 elementary schools in the Midwest. Standardized achievement test scores were once again used as both input and output measures with spring posttest scores regressed onto fall pretest data. The subject area was reading and the reliability reported was .83 for the analysis involving school mean residuals. This compares favorably with the stepped-up results from Dyer, et al. (1969) which were .77, .80, and .86 on the vocabulary, reading and language subtests. Although these reliability coefficients are quite high, it is important to remember that the only factor allowed to vary was the sampling of students from a given school, grade, and year. Therefore, although they could possibly be used to justify computing SEIs for subsamples of students from large schools, they provide no evidence for the consistency of SEIs when these important factors are allowed to vary.

Forsyth (1973) studied the stability issue as it relates to

two successive classes in a particular school. Although differing slightly in some technical details from the studies cited above, the same basic approach of using standardized test results as both inputs and outputs and analyzing school means (n=50) was employed. Outputs were the twelfth grade standardized test scores for two successive classes (graduating in 1968 and 1969) and their test results as ninth graders (in 1965 and 1966) were the inputs. For the nine subtest scores of the Iowa Tests of Educational Development (and the Composite) the correlations among the residuals for the two years ranged from .11 (Quantitative Thinking and Vocabulary) to .50 (Social Studies) with a median of .28. Forsyth considered the consistency of classifications in the five category PI metric and noted that perfect agreement was rare (16 percent to 36 percent). He then argued that for many applications a difference of one category on the PI scale may be sufficiently consistent. Using this criterion, between 62 percent and 88 percent of the schools were "consistently" categorized depending on the subtest under consideration.

In a recent paper Helmstadter and Walton (1985) have presented correlations of SEIs across four elementary grades (third through sixth) and three subject matter areas (math, reading, and language). Based on regression analyses of school means, within grade correlations among the three subject area SEIs were quite large (roughly between .7 and .9). Correlations of SEIs across grades within the same subject area were somewhat smaller, typically between .4 and .6. Although based on large samples (40,000 students per grade in 450 or more schools) the details of the regression analyses are unclear. It is unlikely, however, that the research was based on longitudinal data.

In a study conducted by Matthews, Soder, Ramey and Sanders

(1981) using longitudinally matched data for students attending the Seattle Public Schools, the results were not so positive. Student level residuals using earlier achievement and various SES measures produced SEIs which were quite inconsistent across grades (the grade span was the second to the eighth grade), subject areas (reading and math), and years (1978-79 and 1979-80). The authors discussed but presented no specific results dealing with inconsistencies of positive and negative outliers as a function of subject and grade. As far as differences as a function of year are concerned, however, they noted that in some cases, a school was identified as a positive outlier one year and a negative outlier the next. Year-to-year correlations of SEIs computed at the same grade level ranged from -.24 to .44, none of which were statistically significant because of the small number of schools involved.

Methods and Data Sources

For a number of years, the state of South Carolina has had a policy of statewide testing of students in the majority of the grades in the K-12 grade span. Criterion referenced tests (CRT) used as a part of the Basic Skills Assessment Program (BSAP) are administered each spring to all students in grades 1,2,3,6,8, and 11. Students in grades 4, 7, and 10 are tested, also in the spring, with the Comprehensive Tests of Basic Skills (CTBS). In addition, the Cognitive Skills Assessment Battery (CSAB) is administered at the beginning of the first grade as a readiness test.

The BSAP tests are relatively short and include reading, mathematics, and writing at the higher grade levels. The reading and mathematics subtests contain 36 and 30 multiple choice items respectively. Scale scores are available for the BSAP tests.

This study was limited to the 485 elementary schools in South

Carolina which contain grades one through four (and possibly additional grade levels). Student records for the Spring 1985 testing were matched with the corresponding test records for the previous testing (Spring 1984) with one exception. The first grade BSAP records were matched with the corresponding (Fall 1984) CSAB records. Schools with fewer than 20 matched (and complete) student records at each of the four grade levels were eliminated from consideration reducing the number of schools to 423. In order to obtain stability data comparable to the data presented in the studies cited above, each school-grade sample was split into two random subsamples of equal size. BSAP scale scores in reading and mathematics (grades 1-3), and expanded scale scores for the Total Reading and Total Mathematics subtests of the CTBS (fourth grade) based on the Spring 1985 testing were used as the output variables for each of the four grade cohorts. The "year earlier" BSAP scale scores (for students in grades 2-4 in spring of 1985) or the the CSAB raw score (for 1985 first graders) were considered to be student input variables. Variables representing the percentage of children eligible for free lunches and the percentage eligible for reduced price lunches in 1985 were used as "hard-to-change" variables and students whose records indicated that they were handicapped were eliminated. Regression analyses of the school subsample mean outputs onto the mean inputs and the two lunch percentages were conducted for each of the eight subsamples. Although not precisely in keeping with Dyer's prescription, studentized residuals for reading and mathematics were used as the SEIs.

Reliability coefficients reflecting the consistency of the within-grade subsample SEIs were computed for purposes of comparison with the results cited above. Intraclass correlations (r_I) were obtained to reflect the stability of the subsample SEIs, and were

stepped-up (r_{22}) to reflect the reliability of the results which might be expected for the total sample. Intraclass correlations were selected over the more common Pearson (interclass) correlations since they are more appropriate measures of consistency of the results for the randomly created subsamples. Because of the large sample size, the biased estimator was considered sufficient (see Winer, 1971, p. 287).

SEIs were then recomputed for the total sample. As a matter of interest, these SEIs were correlated with the average of the two subsample SEIs in order to verify that the stepped-up stability coefficients were reasonable in reference to the results based on total samples. To address the main issues in this paper, correlations between the reading and mathematics SEIs within a grade and among the four grade-specific sets of SEIs were obtained. If these results warranted further analysis, the SEIs were dichotomized in order to simulate the selection of schools for an award and the consistency of these decisions was considered using the Kappa coefficient. Finally, similar results were considered in terms of indices obtained by aggregating across the two subject matter areas and the four grade levels.

Results

Results of Preliminary Analyses

A summary of the results of the regression analyses is presented in Table 1. As has been mentioned, these analyses were conducted for all schools in South Carolina with 20 or more useable matched records at the grade level under consideration. To clarify,

---

Insert Table 1 about here

---

schools with grades 1-3 were eligible for inclusion in the analyses

for those three grades but were not included in the final sample of 423 schools. This final sample of 423 schools contained approximately 30,000 first graders with between 20 and 216 matched first-grade records per school; roughly 25,000 second graders with between 22 and 152 per school; approximately 24,000 third graders with 22 to 140 per school; and about 24,500 fourth graders with between 21 and 153 per school.

The results as presented in Table 1 are seen to be quite stable across subsamples but the multiple correlations are somewhat smaller than those reported by Dyer, et al. (1969). It is likely that the primary reason for this finding is that in the Dyer study the output measures were obtained from eighth graders, older students than the first through fourth graders considered here. The data in Table 1 support the common finding that student (and therefore school) achievement is more accurately predicted for older than for younger students. A second explanation for these results might be use of the shorter CRTs at most grade levels.

We also observe that achievement in reading across grade levels is predicted more precisely than achievement in mathematics, a result which tends to be consistent with studies cited above which dealt with children in the early grades (e.g., Webster and Olson, 1984). This interesting finding suggests that more variation in the reading performance of young children can be accounted for in terms of factors such as readiness, previous achievement, and SES than is true of their mathematics performance. A likely causal variable would be the amount of preschool training, possibly at home, and probably concentrated on skills associated with reading. Thus it appears as though there exists more "free" variation in mathematics than reading which suggests that schools could potentially have more of an impact

in this basic skills area. Two somewhat curious findings are: (1) that the percentage of students eligible for reduced price lunches is a predictor of reading but not mathematics achievement, and (2) that previous mathematics performance is not a significant predictor (in the context of the other predictors) of second grade mathematics achievement.

Insert Table 2 about here

In Table 2 the results of correlating the subsample SEIs are presented. The results indicate that performance in mathematics (median stepped-up reliability of .86) was somewhat more stable across subsamples than performance in reading (median reliability of .78). The .78 value for reading compares favorably with the corresponding result obtained by Dyer, et al. (1969) and presented in stepped-up form as .80 by O'Connor (1972). The Dyer results, however, did not indicate that mathematics performance was more stable than reading performance as suggested in Table 2.

Results of Primary Analyses

The results presented above have characterized the consistency of results as they pertain to the sampling variability of student performance within a given grade and subject area. Next we will consider the consistency of SEIs <u>across</u> the two subject areas of reading and mathematics but <u>within</u> grade level. In this case, Pearson correlations are appropriate and are reported in Table 3. The "Total" column in Table 3 refers to the correlations between the reading and math SEIs computed from the total sample. Correlations between these SEI and the average of the two subsample SEIs were all larger than .98, indicating that r_{22} provides a reasonable estimate of the stability of the SEIs based on the total sample. All correlations in

Table 3 are significant and of moderate size, indicating that, within

Insert Table 3 about here

the same grade level, student performance in the two subject areas is reasonably consistent. Although these results are somewhat disquieting, the correlations do not provide a clear picture of the inconsistencies which might arise if the objective were to identify "exceptional" schools based on SEIs for one of the two subject areas. For this purpose, the SEIs were dichotomized to simulate the identification of "exceptional" performance. That is, SEIs in excess of 1.0 (Dyer's criterion for PI=5 was 1.5 but for many applications this would be too selective) were considered exceptional and percentages dealing with decision consistency and coefficient Kappa were obtained. The results are reported in Table 4.

Insert Table 4 about here

The Kappa coefficients range from .52 for first grade to .33 for fourth grade suggesting that decisions based on one or the other of these two important basic skills become less stable as children mature and develop. Since the standard errors are very small for samples of this size, all Kappa coefficients are significant. However, the percentages of inconsistent classifications provide clear evidence that two rather different sets of schools would be identified depending upon whether reading or mathematics were the one basic skills area selected.

It is important to note that the correlations and results on decision consistency above reflect the stability of performance of the same group of students and, therefore, do not reflect inconsistencies which may be introduced if different grades are considered. Table 5

contains the intercorrelations among grade-specific SEIs. These

Insert Table 5 about here

correlations are discouragingly small, the majority not achieving statistical significance at the .05 level. In reading, in particular, there is essentially no relationship between the SEIs for the four grades with the exception that fourth grade SEIs are very moderately related to SEIs reflecting the performance of first and third graders. Although most of the correlations in the mathematics area are large enough to achieve significance, this is little solace if they are considered as parallel forms reliability coefficients. Since the correlations were so small, analyses based on decision consistency were considered unnecessary.

Results Regarding Aggregation

The results presented to this point suggest that SEIs based on reading and mathematics performance of the same student cohort are modestly consistent but that, when the SEIs of students at different grade levels are related, the results border on randomness. Although these findings suggest rather strongly that aggregation of such disparate SEIs will be a fruitless endeavor, for completeness, unweighted average SEIs were computed across the two dimensions of interest in this study. First, the average (AVE) of the reading and math SEIs were obtained at each grade level. The r_{22} indices of these SEIs were very similar to those relating to mathematics only (see Table 2) with the largest difference between the two sets of indices never exceeding .02. This comparability apparently reflects a trade-off between an increase which might be expected for a more comprehensive index and the fact that the reading SEIs are less stable than those reflecting math performance.

Secondly, averages across the four grades for each of the two subject areas and AVE were computed (referred to as composite scores). As might be suspected from the earlier results, this scheme did not produce the increases in stability we normally expect from aggregation. The stability across subject areas of the composite scores was .66 for the total sample, midway between the smallest and largest grade-specific values of .60 and .70 (see Table 3) and Kappa was .42 again representative of the values presented in Table 4. The r_{22} value associated with the composite reading SEI was .80, in the range of the grade-specific values presented in Table 2. The corresponding stability coefficient for mathematics was .90 which is larger than the grade-specific coefficients which ranged from .84 to .87. The stability of the composite based on AVE, which corresponds to an unweighted aggregation across subject areas and grades, was .88 again approximating the stability of the "mathematics only" composite. composite presented.

Discussion and Educational Significance

The approach used in this paper for computing SEIs is clearly not perfect. Arguments concerning the restricted nature of achievement test data and the limited coverage afforded by tests in only two subject areas are clearly valid. Furthermore, no attempt has been made to deal with issues of equity. (The authors acknowledge the importance of assessing school impact on all pupil subpopulations; equity issues were not dealt with in this paper for simplicity alone.) However, the use of a general approach which has been found to have merit by a number of researchers and apply it to large, longitudinally matched samples, appears to be unique. Furthermore, it seems reasonable to presume that in the elementary grades considered here, achievement in reading and mathematics should be priority areas for

all schools. The fact that the BSAP tests were developed based on statewide objectives in reading and mathematics lend further support for this viewpoint and suggests that they should be reasonably "curriculum valid." These results cannot easily be discredited. What, then, are the implications for educational practice?

First, the results should cause "effective schools" researchers to rethink the concept of an effective school. The inconsistency of the results across grades strikes at the very heart of a model which posits school "main effects." In the same vein, Matthews, et al. (1981), discussing the inconsistency of SEIs across two school years (different student cohorts) stated "the low correlations obtained here indicate that high or low performance at a given grade level in a school may have more to do with the characteristics of that particular student cohort than with school effects." (p. 11). Apparently how well a given group of students achieve in a given subject in a given year, when achievement is gauged against potential, is only weakly related to similar measures for other cohorts.

Secondly, the results suggest that school effects, at least at the early grades, may be more or less discernible depending upon the subject area considered. The majority of the inter-grade SEIs in mathematics, although small, were at least larger than chance whereas most of those for reading were not. A suggested explanation of this finding is that young children are more likely to gain knowledge and skills in areas such as reading from sources outside the school than is true for areas such as mathematics. A strategy of identifying effective schools based on mathematics achievement alone in order to achieve more stable SEIs, although psychometrically rational, seems educationally unsound.

The results create a serious problem for those charged with the identification of schools to receive incentive awards based on student achievment. In an attempt to assess schools in a comprehensive fashion, the proposed algorithms usually aggregate grade-subtest SEIs and use the composite index for purposes of award decisions. This is a logically sensible and politically defensible approach. Psychometrically, however, it appears to be analogous to awarding scores to students who randomly responded to a number of test items in that "true score variance" does not seem to manifest itself.

It is possible that the results simply reflect the different goals that school leaders set for themselves each year. Thus, a school might successfully impact on the mathematics performance of low achieving third and fourth graders as intended, but the matrix of SEIs would not demonstrate consistency. This problem appears to be the basis for Rowan's (1985) statement, "The best method of measuring school effectiveness is unknown." (p. 99). For such a model, a school-specific weighting system would be needed if aggregation were to be meaningful.

Common experience suggests that, there are effective and ineffective principals (and other school level staff members) who have an <u>overall</u> positive or negative affect on what happens in a school. Empirical support for this position, at least when effectiveness is measured by residuals from a school level regression analysis, is another matter.

References

Coleman, J. S., Campbell, E. Q., Hobson, C. J., Mood, A. M., Winfield, F. D., & York, R. L. (1966). Equality of educational opportunity. Washington, D. C.: U. S. Department of Health, Education and Welfare.

Dyer, H. S. (1970). Toward objective criteria of professional accountability in the schools of New York City. Phi Delta Kappan, 52, 206-211.

Dyer, H. S., Linn, R. L., & Patton, M. J. (1969). A comparison of four methods of obtaining discrepancy scores based on observed and predicted school system means on achievement tests. American Educational Research Journal, 6, 591-605.

Forsyth, R. A. (1973). Some empirical results related to the stability of performance indicators in Dyer's student change model of an educational system. Journal of Educational Measurement, 10, 7-12.

Frechtling, J. A. (1982). Alternative methods for determining effectiveness: Convergence and divergence. Paper presented at the annual meeting of the American Educational Research Association, New York.

Frederick, J. M. & Clauset, K. H. (1985). A comparison of the major algorithms for measuring school effectiveness. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Helmstadter, G. C. & Walton, M. A. (1985). The generalizability of residual indexes of effective schooling.

Hilton, T. L., and Patrick, C. (1970). Cross-sectional versus longitudinal data: An empirical comparison of mean differences in academic growth. Journal of Educational Measurement, 7, 15-24.

Marco, G. (1974). A comparison of selected school effectiveness measures based on longitudinal data. Journal of Educational Measurement, 11, 225-234.

Matthews, T. A. & Walton, G. C. (1981). Use of district test scores to compare the academic effectiveness of schools. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles.

O'Connor, E. F. (1972). Extending classical test theory to the measurement of change. Review of Educational Research, 42, 73-97.

Ralph, J. H., and Fennessey, J. (1983). Science or reform: Some questions about the effective schools model. Phi Delta Kappan, 64, 689-702.

Rowan, B. (1985). The assessment of school effectiveness. In Kyle, R. M. J. (Ed.). Reaching for excellence: An effective schools sourcebook. N. I. E.: Washington, D. C.

Rowan, B., Bossart, S. T., and Dwyer, D. C. (1983). Research on effective schools: A cautionary note. Educational Researcher, 12, 24-31.

Winer, B. J. (1971). Statistical principles in experimental design. New York: McGraw-Hill.

Wynne, E. A. (1984). School award programs: Evaluation as a component of incentive systems. Educational Evaluation and Policy Analysis, 6, 85-93.

Table 1

Significant Predictors and Squared Multiple Rs
By Output Variable and Grade

Grade	N	Output	Significant Predictors	R^2 Sub1	R^2 Sub2	R^2 Tot
1	533	BSAP-R	CSAB LUNCHF	.45	.46	.48
		BSAP-M	CSAB LUNCHF	.30	.33	.34
2	519	BSAP-R	BSAP-R LUNCHF LUNCHR	.64	.65	.68
		BSAP-M	BSAP-R LUNCHF	.44	.43	.46
3	523	BSAP-R	BSAP-R LUNCHF LUNCHR	.63	.62	.66
		BSAP-M	BSAP-R BSAP-M LUNCHF	.36	.29	.34
4	508	CTBS-R	BSAP-R LUNCHF LUNCHR	.72	.74	.76
		CTBS-M	BSAP-R BSAP-M LUNCHF	.47	.51	.50

Note: To be included as a "significant predictor", a regression coefficient was significant ($p < .05$) for all three analyses. This excluded only two cases in which a predictor was significant for one of the two subsamples.

Table 2
Intraclass Correlations and Stepped-Up Reliabilities
Measuring Consistency of Subsample SEIs
By Subject Area and Grade

	Reading		Math	
Grade	r1	r22	r1	r22
1	.76	.86	.77	.87
2	.56	.71	.73	.84
3	.63	.77	.76	.86
4	.65	.79	.76	.86
Median(1-4)	.64	.78	.76	.86

Note: Due to rounding, some of the results do not precisely agree with the Spearman-Brown formula.

BEST COPY AVAILABLE

Table 3
Pearson Correlations Between Reading and Mathematics SEIs
For Each Subsample and the Total Sample

Grade	Subsample 1	Subsample 2	Total
1	.65	.69	.70
2	.49	.59	.60
3	.55	.54	.60
4	.60	.61	.63

Table 4
Decision Consistency By Grade
For Reading and Mathematics SEIs
For Total Sample

	Percentages			
Grade	--	+-/-+	++	Kappa
1	79.7	11.8	8.5	.52
2	78.3	15.6	6.1	.53
3	80.4	12.5	7.1	.46
4	77.8	16.3	5.9	.33

Note: A "+" sign indicates "exceptional" according to the definition in the text.

Table 5
Pearson Correlations Among Grade-Specific SEIs
By Subject Area

	Reading			Mathematics		
Grades	Sub1	Sub2	Total	Sub1	Sub2	Total
1 & 2	-.02	-.01	.02	.12*	.15**	.16**
1 & 3	.06	.05	.06	.14**	.08	.14**
1 & 4	.13**	.06	.11*	.08	.04	.08
2 & 3	.09	.00	.07	.08	.17**	.17**
2 & 4	.06	.03	.04	.06	.10*	.11*
3 & 4	.06	.15**	.14**	.09	.06	.11*
Median	.06	.04	.06	.09	.09	.13**

Note: * $p < .05$; ** $p < .01$.